



Article

# SMD-Net: Selective and Multiway Differential Perception-Enhanced Network for Early Multimodal Rumor Detection

Zhengnan Qiao , Zhekang Yang and Xianguo Zhang \*

College of Computer Science, Inner Mongolia University, Hohhot 010021, China;  
qiaozhengnan@mail.imu.edu.cn (Z.Q.); 32409031@mail.imu.edu.cn (Z.Y.)

\* Correspondence: cszxcg@imu.edu.cn

## Abstract

The rapid dissemination of rumors on social media at their early stages poses significant threats to public safety and social stability. While response-based methods usually depend on user comments and reposts and therefore suffer from inherent latency, existing content-based methods still struggle to extract discriminative evidence from noisy short texts and subtle visual inconsistencies under zero-response conditions. To address this issue, we propose SMD-Net, a multimodal framework for early zero-response rumor detection. In the textual branch, a selective state-space encoder is used to model fragmented and noisy posts. In the visual branch, an enhanced TransXNet backbone is designed to improve the representation of fine-grained suspicious patterns and cross-layer feature interactions. An adaptive gated fusion module is further introduced to integrate textual and visual features for final prediction. Experiments on the Weibo and PHEME datasets show that SMD-Net outperforms the compared content-based baselines, achieving 92.60% accuracy on Weibo and 90.27% accuracy on PHEME under the strict zero-response setting. These results suggest that the proposed framework provides an effective solution for early multimodal rumor detection when propagation-based evidence is unavailable.

**Keywords:** early rumor detection; multimodal rumor detection; zero-response scenario; Mamba; differential attention; multimodal fusion; gated fusion

## 1. Introduction

The widespread adoption of social media platforms such as Weibo and Twitter has significantly accelerated the dissemination of misinformation and rumors. Rumors in social media are typically defined as unverified claims that propagate rapidly across networks by exploiting cognitive biases. Empirical evidence indicates that the impact of misinformation extends beyond online environments, affecting financial markets, public health systems, and political processes. Consequently, the development of automated early rumor detection methods from large-scale, fragmented data streams has become a critical requirement for maintaining information integrity and social stability. However, accurately identifying rumors at an early stage remains challenging due to the limited availability of reliable contextual signals and the high level of noise in raw multimodal content [1–4].

Early studies on automated rumor detection primarily relied on expert-curated knowledge bases or external authoritative evidence. However, due to the rapid emergence of social media content, such approaches often suffer from limited timeliness [5,6]. Subsequent research shifted toward exploiting collective intelligence from social media by utilizing



Academic Editor: Alexey Karpov

Received: 10 June 2026

Revised: 3 July 2026

Accepted: 3 July 2026

Published: 10 July 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

user comments and repost patterns to construct response-based multimodal detection approaches. Although response features improve detection accuracy, these methods require waiting for feedback to accumulate at scale, thereby introducing significant time latency and limiting intervention during the early propagation stage of rumors [7,8].

In contrast, content-based approaches have attracted increasing attention because they enable early rumor detection without relying on propagation signals. In this setting, a multimodal post is classified as rumor or non-rumor using only the source text and its associated image, without access to user comments or reposts [9,10]. However, existing content-based methods still face difficulties in zero-response scenarios, where prediction must be made solely from limited and noisy multimodal evidence.

From the visual perspective, many approaches rely on backbone networks such as VGG [11] or ResNet [6,12] to extract global representations. Such representations are often insufficient for preserving subtle local inconsistencies, and global pooling may suppress high-frequency details that are useful for identifying suspicious patterns, such as boundary discontinuities, illumination inconsistencies, and abnormal edge structures [9], as illustrated in Figure 1. From the textual perspective, social media posts are typically short, noisy, and fragmented. Without external context, conventional sequence models often struggle to retain sufficiently informative semantics under low signal-to-noise conditions [13]. In addition, although some studies have explored cross-modal associations through attention mechanisms or multimodal fusion networks [14–16], their ability to capture subtle semantic inconsistency is still limited when intra-modal evidence is not adequately represented. Therefore, early multimodal rumor detection under zero-response conditions still requires more effective extraction of textual semantics, fine-grained visual cues, and cross-modal discrepancy signals.



**Figure 1.** Illustration of subtle visual inconsistencies in a rumor image. The highlighted region is enlarged to reveal weak boundary continuity and texture irregularities, which are difficult to capture using global feature representations.

To address these limitations, this study proposes SMD-Net, a Selective and Multiway Differential Perception-Enhanced Network for early multimodal rumor detection under zero-response conditions. The overall design is motivated by three complementary requirements in this setting: extracting informative semantics from noisy short texts, preserving fine-grained suspicious evidence from images, and integrating the two modalities in a way that reflects both agreement and discrepancy. Accordingly, the textual branch adopts a selective state-space modeling mechanism based on Mamba [17] to encode fragmented textual inputs. The visual branch builds on an enhanced TransXNet backbone [18], which is further enhanced with Differential Attention (DA) [19] and Multiway Dynamic Dense Connection (MUDD) [20] to improve the representation of local irregularities and cross-layer feature interactions. To strengthen cross-modal interaction, an Adaptive Gated Fusion (AGF) module is introduced to combine textual and visual features in a sample-adaptive manner.

Through this design, SMD-Net aims to improve early rumor detection performance using only raw multimodal content, without relying on comments, reposts, or other propagation-based feedback.

The main contributions of this study are summarized as follows:

- (1) We propose SMD-Net, a multimodal framework for early rumor detection under strict zero-response conditions. The framework is designed to address three coupled challenges in this setting: noisy and fragmented textual semantics, subtle visual inconsistency cues, and effective cross-modal integration without propagation-based evidence.
- (2) A selective state-space textual encoder and an enhanced TransXNet-based visual encoder are jointly developed to improve unimodal evidence extraction. In particular, DA and MUDD are incorporated into the visual branch to strengthen the representation of local irregularities and cross-layer visual evidence.
- (3) An adaptive gated fusion module is introduced to integrate textual and visual representations by considering both interaction and discrepancy information between modalities.
- (4) Extensive experiments on the Weibo and PHEME datasets under strict zero-response settings show that SMD-Net consistently outperforms the compared content-based baselines, achieving 92.60% accuracy on Weibo and 90.27% accuracy on PHEME.

## 2. Related Work

The widespread adoption of mobile Internet technologies has made social media platforms primary channels for information dissemination and public opinion expression. However, the low barrier to content creation has led to a high volume of fragmented claims with uncertain veracity in heterogeneous information streams. Rumors, defined as unverified information that spreads rapidly through social networks, pose substantial risks to social stability and public interests. Rumor detection seeks to determine the veracity of such information by exploiting multimodal evidence and has become an active research area. Existing approaches can be categorized based on the availability of verification evidence into three paradigms: (1) fact-checking-based methods, which rely on external knowledge sources; (2) response-based methods, which exploit user interaction and propagation signals; and (3) content-based methods, which perform detection using only the original multimodal content under zero-response conditions. Among these paradigms, content-based methods are particularly suitable for early-stage rumor detection, as they do not rely on external evidence or user interactions.

Fact-checking-based methods verify claims by leveraging external authoritative knowledge sources, providing strong logical support for rumor detection. Early studies aligned textual content with structured knowledge bases, such as knowledge graphs [21], to validate claims through entity relations and topological consistency. With the advancement of large language models (LLMs), retrieval-augmented generation (RAG) has become a prevalent paradigm. For example, WKGFC [22] dynamically retrieves real-time evidence from external sources, while VeraCT [23] enhances interpretability by explicitly modeling the verification process. Despite their effectiveness, fact-checking-based methods suffer from a fundamental limitation referred to as knowledge lag. During the early stages of emerging events, external knowledge bases often lack sufficient evidence to support reliable verification [24]. Even LLM-integrated retrieval systems are constrained by pre-training data and remain susceptible to hallucinations on novel events. This limitation restricts the applicability of fact-checking-based methods in early rumor detection scenarios and highlights the necessity of content-based approaches under zero-response conditions.

Response-based methods assess information credibility by exploiting user-generated signals, including comments, reposts, and propagation structures. Early studies extracted features from comment text, user profiles, and propagation structures for classification [25–27].

Subsequent research modeled temporal patterns in comment sequences using recurrent neural networks [28] or employed tree-structured networks to capture hierarchical topologies in conversation trees [29,30]. More recently, graph neural networks (GNNs) have become the dominant framework for modeling complex propagation structures. For instance, FGDGNN [31] constructs dynamic propagation graphs to capture temporal evolution, while HAGNN [32] introduces hierarchical word-text graphs to enhance the representation of fragmented content. To further improve reasoning capability, recent methods integrate GNNs with large language models. SePro [33] performs cooperative reasoning by extracting key evidence from propagation structures and generating explanatory outputs, while ISMAF [34] incorporates social context as an auxiliary modality to model collective user behavior signals.

Despite their high benchmark accuracy, response-based methods suffer from the cold-start problem due to sparse early-stage user feedback. This dependency introduces inherent latency, as reliable patterns only emerge after widespread dissemination. Furthermore, these methods are constrained by limited platform data access and incur substantial computational costs when modeling large-scale interaction graphs. These limitations collectively hinder their effectiveness for real-time early warning, particularly in zero-response scenarios.

Content-based methods assess the credibility of information using only the original post content, without relying on external knowledge or user interactions. This paradigm is particularly suitable for early-stage rumor detection under zero-response conditions. Early studies focused on handcrafted features, including linguistic patterns and visual artifacts, to capture explicit inconsistencies within multimodal content [35]. With the advancement of deep learning, multimodal fusion has become the dominant approach. For example, Qian et al. [35] proposed a hierarchical attention network to learn joint representations across modalities, while Qi et al. [36] introduced fine-grained alignment mechanisms to model entity-level inconsistencies. To improve generalization across events, EANN [37] and MDDA [38] employed adversarial learning and domain adaptation to extract event-invariant features. More recent studies further enhance semantic modeling from multiple perspectives. MSynFD [39] leverages syntactic dependency graphs to capture structural characteristics of deceptive content, while Dm-inter [40] utilizes sequence-to-sequence models to infer underlying generation intentions. NSLM [41] introduces neural-symbolic reasoning to model latent inconsistency patterns, including visual manipulation and cross-modal conflicts.

Despite these advances, content-based approaches remain constrained in zero-response settings: (1) intra-modal feature extraction is insufficient. Existing visual backbones predominantly rely on global representations, which are not effective in preserving fine-grained local inconsistencies, while conventional text encoders struggle to extract discriminative semantics from short and noisy sequences; (2) cross-modal interaction is typically modeled at a coarse level, which limits the ability to capture subtle semantic inconsistencies between textual content and visual evidence; and (3) existing methods rarely model both microscopic tampering artifacts and high-level logical conflicts in a unified manner, resulting in limited discriminative capability when handling complex and ambiguous rumors.

To address the identified limitations in intra-modal feature extraction and cross-modal inconsistency modeling, this paper proposes the Selective and Multiway Differential Perception-Enhanced Network (SMD-Net) for early multimodal rumor detection under zero-response conditions. The framework employs a Mamba-based selective state-space model in the textual branch to capture global semantic dependencies from noisy and fragmented sequences. In the visual branch, an enhanced TransXNet backbone integrates

a Multiway Dynamic Dense Connection (MUDD) module and a Differential Attention mechanism to improve the extraction of microscopic tampering artifacts. An Adaptive Gated Fusion (AGF) module is subsequently introduced to explicitly model fine-grained consistency and potential semantic conflicts across modalities. Experimental results on the Weibo and PHEME datasets under zero-response settings show that the proposed method consistently outperforms existing content-based approaches and achieves competitive or better performance compared with representative response-based methods.

### 3. Materials and Methods

#### 3.1. Problem Formulation

This study investigates early multimodal rumor detection on social media, where rumors propagate rapidly at the initial stage and may cause a significant impact. Unlike response-based methods that rely on user interactions such as comments, reposts, or stance signals, the proposed framework operates under a zero-response setting and performs classification using only the original post content.

For each sample, the multimodal input is defined as:

$$M = \{T, I\} \quad (1)$$

where  $T = \{w_1, w_2, \dots, w_n\}$  denotes the textual sequence consisting of  $n$  tokens, and  $I \in \mathbb{R}^{H \times W \times C}$  represents the associated image, with  $H$ ,  $W$ , and  $C$  denoting height, width, and channel dimensions, respectively.

The objective is to learn a mapping function:

$$F : \{T, I\} \rightarrow y \quad (2)$$

where  $y \in \{0, 1\}$  is the class label, with  $y = 1$  indicating rumor and  $y = 0$  indicating non-rumor, and  $F(\cdot)$  denotes the proposed multimodal classification model.

The primary challenge lies in the early-stage setting, where textual content is noisy and semantically incomplete, while images may contain subtle manipulations or cross-modal inconsistencies. Reliable discriminative evidence must therefore be extracted from limited and heterogeneous information without any propagation context.

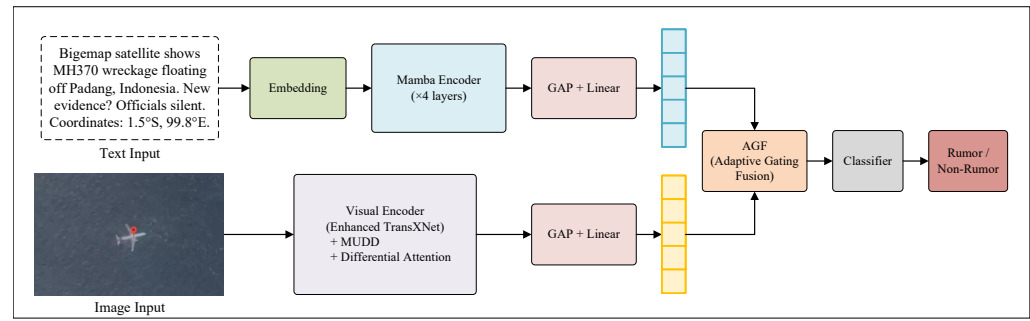
#### 3.2. Methodology

The overall architecture of SMD-Net is illustrated in Figure 2. The model performs early multimodal rumor detection under zero-response conditions using only the original post content. SMD-Net consists of three stages:

**Unimodal Feature Encoding.** The textual input  $T$  and the associated image  $I$  are mapped into initial representations. The text is converted into a sequence of token embeddings, while the image is encoded by the visual backbone to obtain a representation for subsequent multimodal processing.

**Evidence Refinement and Enhancement.** The two modalities are processed separately to extract discriminative features. The textual branch employs a selective state-space model to model long-range dependencies in textual sequences. The visual branch is built upon an enhanced TransXNet backbone, where differential attention (DA) and Multiway Dynamic Dense Connection (MUDD) are introduced to model local irregularities and cross-layer feature interactions.

**Adaptive Cross-modal Fusion.** The modality-specific features are integrated through an adaptive gated fusion module, producing a joint representation  $z_{fusion}$  for final prediction.



**Figure 2.** Overall architecture of SMD-Net for early multimodal rumor detection under zero-response conditions.

### 3.2.1. Multimodal Feature Encoding

Given a multimodal sample  $M = \{T, I\}$ , where  $T = \{w_1, w_2, \dots, w_L\}$  denotes the tokenized textual sequence and  $I \in \mathbb{R}^{H \times W \times C}$  denotes the associated image, the encoding stage maps each modality into an initial latent representation. For implementation, the textual sequence is truncated or padded to a fixed length  $L$ .

**Textual Feature Initialization.** Let  $\mathcal{V}$  denote the vocabulary built from the training corpus, with size  $N_v = |\mathcal{V}|$ . Each token in  $T$  is mapped to an index in  $\mathcal{V}$  and then converted into a dense vector through a trainable embedding matrix  $\mathbf{W}_{emb} \in \mathbb{R}^{N_v \times D_t}$ . The initial textual representation is formulated as:

$$\mathbf{X}_T^{(0)} = \text{Embed}(T; \mathbf{W}_{emb}), \quad (3)$$

where  $\mathbf{X}_T^{(0)} \in \mathbb{R}^{L \times D_t}$  denotes the embedded textual sequence.

**Visual Feature Encoding.** For the visual modality, the input image  $I$  is processed by the enhanced visual backbone implemented with TransXNet to obtain a global visual representation:

$$\mathbf{f}_I^{(0)} = \text{VisualEncoder}(I), \quad (4)$$

where  $\mathbf{f}_I^{(0)} \in \mathbb{R}^{D_v}$  denotes the visual feature extracted from the final stage of the backbone. To align the visual feature with the multimodal fusion space, a linear projection is applied:

$$\mathbf{f}_I = \mathbf{W}_I \mathbf{f}_I^{(0)} + \mathbf{b}_I, \quad (5)$$

where  $\mathbf{W}_I \in \mathbb{R}^{D \times D_v}$  and  $\mathbf{b}_I \in \mathbb{R}^D$  are trainable parameters.

### 3.2.2. Textual Evidence Refinement Module

Early social media posts are typically short, noisy, and semantically fragmented, which makes discriminative semantic extraction from raw text challenging. To encode token dependencies under this setting, the textual branch adopts a selective state-space model Mamba [17] as the sequence encoder. In this scenario, not all tokens carry equal discriminative value. The core mechanism of Mamba—an input-dependent selective state space—allows the sequence model to adaptively regulate state transitions based on the current input. This property enables the encoder to suppress responses to noise tokens while retaining informative semantics, which directly addresses the low signal-to-noise condition of early social media posts.

Given the embedded textual sequence  $\mathbf{X}_T^{(0)} \in \mathbb{R}^{L \times D}$ , the textual branch applies a stacked selective state-space encoder to refine the initial token representations:

$$\mathbf{H}'_T = \text{SSMEncoder}(\mathbf{X}_T^{(0)}), \quad (6)$$

The encoder consists of multiple residual blocks. For the  $l$ -th block:

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \mathcal{F}(\text{RMSNorm}(\mathbf{H}^{(l)})), \quad (7)$$

where  $\mathcal{F}(\cdot)$  denotes the selective state-space transformation, and  $\text{RMSNorm}(\cdot)$  denotes root mean square normalization. Within each block, the input is first projected into a main branch and a gating branch. The main branch is processed by a depthwise one-dimensional convolution followed by an input-dependent state-space update, while the auxiliary branch provides element-wise gating to modulate the transformed features. After sequence refinement, global average pooling is applied:

$$\mathbf{f}_T^{(0)} = \frac{1}{L} \sum_{i=1}^L \mathbf{H}_T'[i, :], \quad (8)$$

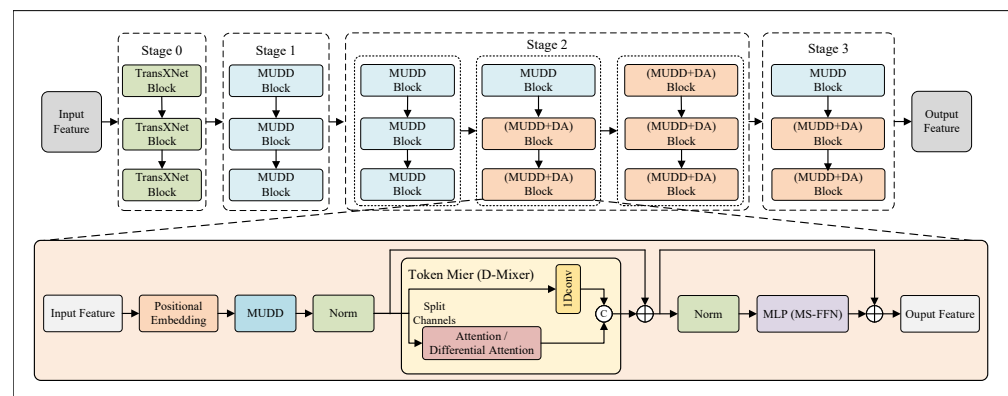
The pooled feature is then projected into the shared multimodal representation space:

$$\mathbf{f}_T = \mathbf{W}_T \mathbf{f}_T^{(0)} + \mathbf{b}_T, \quad (9)$$

### 3.2.3. Visual Evidence Enhancement

The visual branch aims to extract discriminative evidence from rumor images by modeling local inconsistencies and hierarchical feature interactions. To this end, the model employs an enhanced TransXNet backbone to progressively refine visual representations under the zero-response setting. Starting from the patch-embedded visual feature map, the encoder performs hierarchical representation learning across four stages and produces a global visual representation for subsequent multimodal processing.

**Enhanced TransXNet Backbone.** As shown in Figure 3, the visual encoder adopts a four-stage TransXNet backbone [18]. The input image is first converted into an initial feature map by patch embedding. The feature map is then processed by stacked blocks across four stages, with spatial resolution decreasing and channel dimension increasing stage by stage. Each block contains positional encoding, hybrid token mixing, and a feed-forward transformation with residual connections.



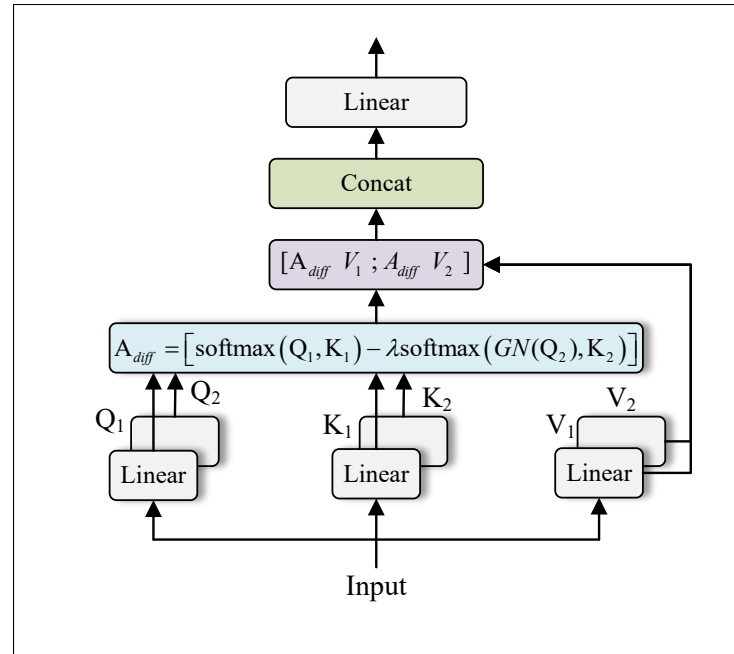
**Figure 3.** Detailed architecture of the enhanced visual encoder in SMD-Net.

Two enhancement modules are introduced at different stages of the backbone. Multiway Dynamic Dense Connection (MUDD) [20] is introduced in the deeper stages to aggregate historical features and strengthen cross-layer feature propagation. Differential Attention (DA) [19] is also used in deeper stages, replacing standard global attention to highlight subtle local inconsistencies in high-level representations.

After the final stage, the output feature is aggregated into a global visual representation for subsequent multimodal processing.

### Differential Attention.

As shown in Figure 4, Differential Attention (DA) is introduced in the deeper layers of the visual encoder to strengthen the sensitivity of the image modality to subtle tampering traces. Unlike global representations that may smooth weak local anomalies, DA enhances fine-grained irregularities, such as splice boundaries, over-sharpened edges, and abnormal blur patterns.



**Figure 4.** Structure of the differential attention (DA) module.

Given an input feature map  $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ , the module first generates the query, key, and value representations, denoted by  $\mathbf{Q}$ ,  $\mathbf{K}$ , and  $\mathbf{V}$ , where the spatial dimension is flattened into  $N = H \times W$  tokens. The channels of  $\mathbf{Q}$ ,  $\mathbf{K}$ , and  $\mathbf{V}$  are evenly divided into two groups, yielding  $\{\mathbf{Q}_1, \mathbf{Q}_2\}$ ,  $\{\mathbf{K}_1, \mathbf{K}_2\}$ , and  $\{\mathbf{V}_1, \mathbf{V}_2\}$ . Two attention maps are then computed as:

$$\mathbf{A}_1 = \text{Softmax}\left(\frac{\mathbf{Q}_1^\top \mathbf{K}_1}{\sqrt{d_h}}\right), \quad \mathbf{A}_2 = \text{Softmax}\left(\frac{\text{GN}(\mathbf{Q}_2)^\top \mathbf{K}_2}{\sqrt{d_h}}\right), \quad (10)$$

where  $d_h$  denotes the channel dimension of each attention head, and  $\text{GN}(\cdot)$  denotes group normalization. A learnable balancing parameter  $\lambda$  is introduced to construct the differential attention map, which follows the original differential attention formulation [19]. Here,  $\lambda$  is initialized as  $\lambda_{\text{init}} = 0.8$ , and the training procedure allows it to converge to a value that balances the suppression of shared texture responses against the preservation of distinctive patterns:

$$\mathbf{A}_{\text{diff}} = \mathbf{A}_1 - \lambda \mathbf{A}_2, \quad (11)$$

The resulting map is applied to the value features, followed by channel fusion, to obtain the refined output:

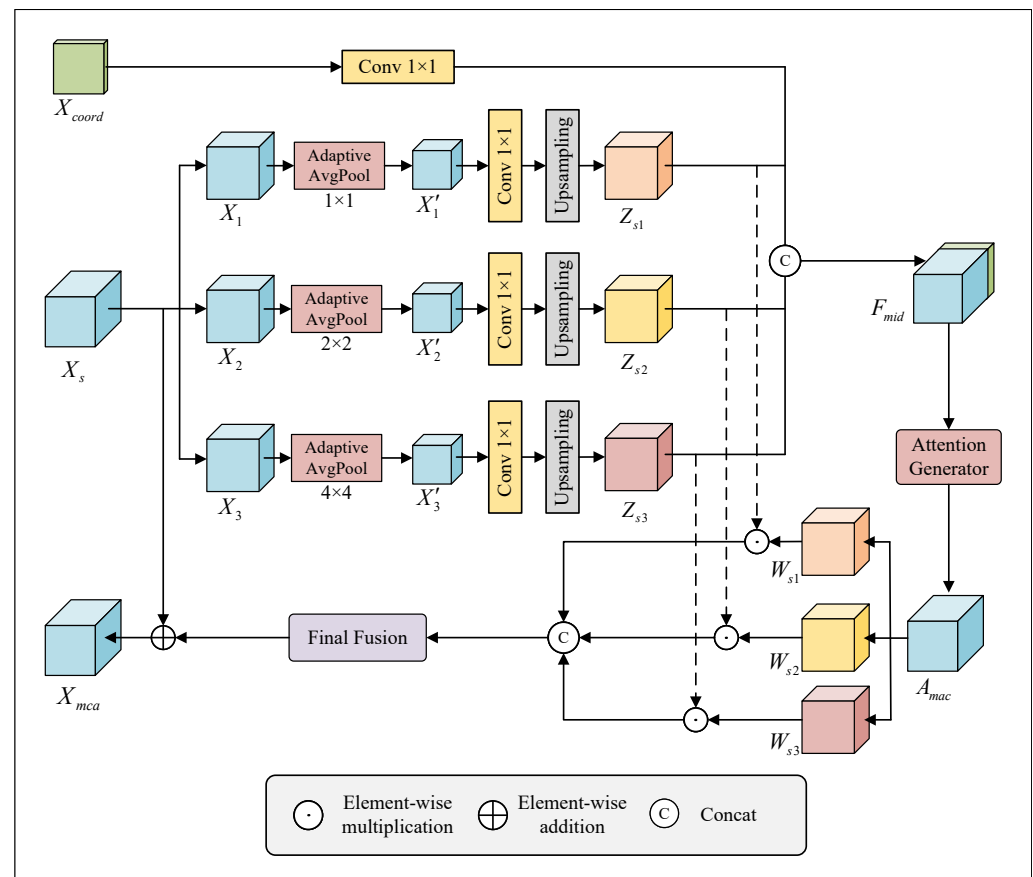
$$\mathbf{X}_{\text{out}} = \mathcal{R}\left(\text{Conv}_{1 \times 1}\left([\mathbf{A}_{\text{diff}} \mathbf{V}_1^\top, \mathbf{A}_{\text{diff}} \mathbf{V}_2^\top]^\top\right)\right), \quad (12)$$

where  $[\cdot]$  denotes channel-wise concatenation and  $\mathcal{R}(\cdot)$  reshapes the token representation back to the spatial feature map.

By subtracting the secondary attention response from the primary one, DA suppresses shared texture responses and highlights locally distinctive patterns, thereby improving the representation of microscopic tampering cues in the visual modality.

### Multiway Dynamic Dense Connection.

In deep visual encoding, fine-grained cues may be weakened by successive feature transformations. To mitigate this effect, a Multiway Dynamic Dense Connection (MUDD) module is employed, as shown in Figure 5, to aggregate historical features from preceding blocks and improve feature interaction across the visual branch.



**Figure 5.** Structure of the Multiway Dynamic Dense Connection (MUDD) module.

Given the current feature map at the  $l$ -th block, denoted by  $\mathbf{X}_l \in \mathbb{R}^{C \times H_l \times W_l}$ , MUDD first generates path-wise dynamic aggregation weights from the global context of the current feature:

$$\mathbf{W}_l = \text{Softmax}(\psi(\text{GAP}(\mathbf{X}_l))), \quad (13)$$

where  $\psi(\cdot)$  denotes a lightweight projection network,  $\mathbf{W}_l \in \mathbb{R}^{P \times m_l}$  is the dynamic weight matrix,  $P$  is the number of paths, and  $m_l$  is the number of available historical features.

Next,  $\mathbf{X}_l$  is evenly split along the channel dimension into  $P$  sub-features:

$$\mathbf{X}_l = \{\mathbf{X}_l^{(1)}, \mathbf{X}_l^{(2)}, \dots, \mathbf{X}_l^{(P)}\}, \quad (14)$$

For each path  $p \in \{1, \dots, P\}$ , a path-specific transformation  $\phi^{(p)}(\cdot)$  is applied to the current feature, and the historical features from preceding blocks are spatially aligned to the current resolution and aggregated with dynamic weights:

$$\mathbf{Y}_l^{(p)} = \phi^{(p)}(\mathbf{X}_l^{(p)}) + \sum_{i=1}^{m_l} \mathbf{W}_l^{(p,i)} \tilde{\mathbf{X}}_{l-i}^{(p)}, \quad (15)$$

where  $\tilde{\mathbf{X}}_{l-i}^{(p)}$  denotes the aligned path-wise feature from the  $i$ -th preceding block.

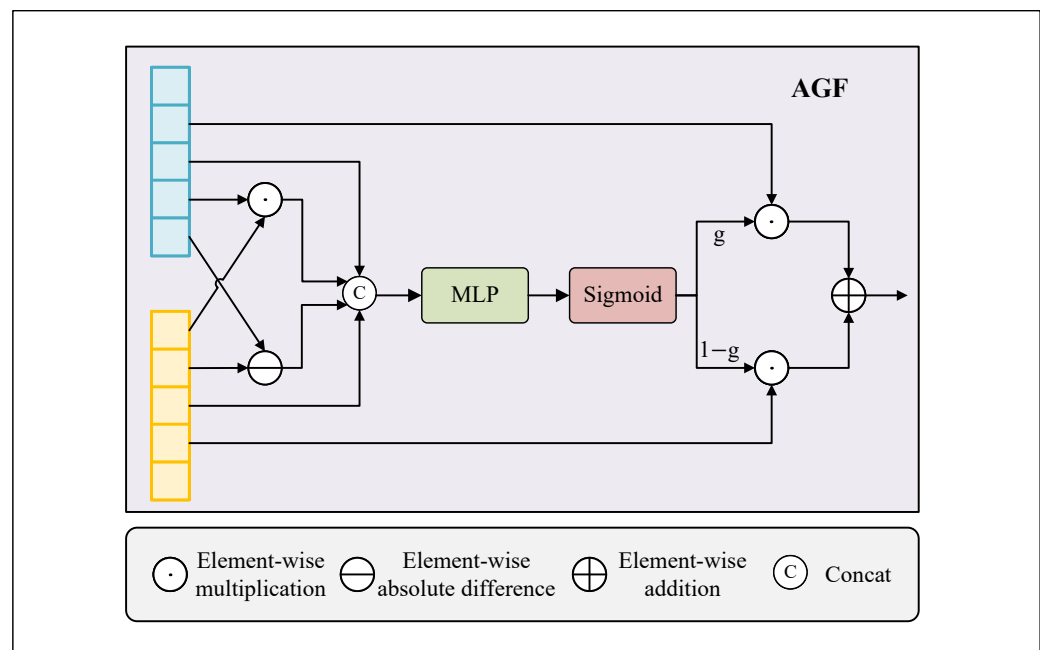
Finally, the enhanced outputs of all paths are concatenated:

$$\mathbf{X}_{l+1} = [\mathbf{Y}_l^{(1)}, \mathbf{Y}_l^{(2)}, \dots, \mathbf{Y}_l^{(P)}], \quad (16)$$

This design helps retain fine-grained cues in deep layers and improves cross-block feature reuse.

### 3.2.4. Adaptive Multimodal Feature Fusion

After modality-specific evidence refinement, the global textual feature  $\mathbf{f}_T \in \mathbb{R}^d$  and the global visual feature  $\mathbf{f}_I \in \mathbb{R}^d$  are integrated by an Adaptive Gated Fusion (AGF) module as shown in Figure 6. Instead of directly concatenating the two modalities, AGF learns sample-specific fusion weights to balance their relative contributions while preserving both cross-modal agreement and discrepancy cues.



**Figure 6.** Structure of the adaptive gated fusion (AGF) module.

First, a joint interaction vector is constructed as:

$$\mathbf{c} = [\mathbf{f}_I; \mathbf{f}_T; \mathbf{f}_I \odot \mathbf{f}_T; |\mathbf{f}_I - \mathbf{f}_T|] \in \mathbb{R}^{4d}, \quad (17)$$

where  $[\cdot; \cdot]$  denotes vector concatenation and  $\odot$  denotes element-wise multiplication. Here,  $\mathbf{f}_I \odot \mathbf{f}_T$  captures cross-modal agreement, while  $|\mathbf{f}_I - \mathbf{f}_T|$  encodes discrepancy between the two modalities.

The interaction vector is then fed into a two-layer multilayer perceptron to generate an adaptive gate:

$$\mathbf{g} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{c} + \mathbf{b}_1) + \mathbf{b}_2), \quad (18)$$

where  $\mathbf{g} \in (0, 1)^d$ ,  $\delta(\cdot)$  denotes the ReLU activation, and  $\sigma(\cdot)$  denotes the Sigmoid function.

The fused representation is defined as:

$$\mathbf{f}_{\text{fused}} = \mathbf{g} \odot \mathbf{f}_I + (1 - \mathbf{g}) \odot \mathbf{f}_T, \quad (19)$$

In this way, AGF adaptively adjusts the contribution of each modality for each sample and produces a compact multimodal representation for final rumor classification.

### 3.2.5. Objective Function

Given the fused multimodal representation  $\mathbf{f}_{\text{fused}}$ , the classifier outputs a logit vector  $\mathbf{z}_i \in \mathbb{R}^2$  for the  $i$ -th sample. The model is optimized using the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(z_{i,y_i})}{\sum_{c=1}^2 \exp(z_{i,c})}, \quad (20)$$

where  $y_i \in \{0, 1\}$  denotes the ground-truth label of the  $i$ -th sample.

## 4. Results and Discussion

### 4.1. Datasets and Zero-Response Setup

We evaluate the proposed method on two public multimodal rumor detection benchmarks, Weibo [42] and PHEME [43]. The Weibo dataset is collected from Sina Weibo and contains source posts paired with associated images and corresponding binary veracity labels. The PHEME dataset is derived from Twitter and likewise provides source posts, images, and binary labels for rumor detection.

To match the early zero-response setting, only the original post text and its associated image are retained for each sample, while all subsequent propagation signals, such as comments, reposts, and other interaction information, are excluded.

### 4.2. Baselines

We compare SMD-Net with two groups of baselines: content-based methods and response-based methods. Content-based methods make predictions using only the original post content and therefore provide the most direct comparison for the strict zero-response setting considered in this study. By contrast, response-based methods additionally incorporate user comments, stance signals, propagation structures, or social-context information. These methods are not directly comparable to zero-response models, but they are included as broader references to contextualize the performance of SMD-Net.

#### Content-based baselines:

- (a) EANN [37]—uses event-adversarial learning to extract event-invariant multimodal features.
- (b) SAFE [44]—a multimodal similarity-based framework for modeling cross-modal consistency.
- (c) DMFL [45]—employs a dual-chain architecture to capture semantic and logical relations across modalities.
- (d) SemSim [46]—diffusion-based semantic similarity combined with multimodal interaction modeling.

#### Response-based baselines:

- (e) QSAN [47]—jointly models rumor content and user stance signals.
- (f) EBGCN [48]—an edge-enhanced Bayesian graph convolutional network for rumor propagation modeling.
- (g) MFAN [49]—integrates textual, visual, and social graph information through hierarchical multimodal attention.
- (h) GINMCL [50]—graph representation learning combined with contrastive multimodal consistency learning over text, images, and social context.
- (i) CLFFRD [51]—integrates curriculum learning with fine-grained multimodal fusion to progressively train on text, image, and multimodal difficulty levels while using linear interpolation for data augmentation.
- (j) DEEP [52]—uses large language models to extract contrasting evidence from multimodal content and user responses.

#### 4.3. Implementation Details

The proposed model is implemented in PyTorch version 2.7.0 and trained on a server equipped with an NVIDIA RTX 4090 GPU. For the textual branch, a 4-layer Mamba encoder is adopted with a model dimension of 128 and a state dimension of 16. Each post is tokenized and truncated or padded to 128 tokens, and the vocabulary size is set to 10,000. For the visual branch, TransXNet is used as the backbone. Input images are resized to  $256 \times 256$ . During training, data augmentation includes random flipping, rotation, and color jitter, while only center cropping and normalization are applied during evaluation.

The model is optimized using AdamW with an initial learning rate of  $3 \times 10^{-5}$  and a weight decay of  $1 \times 10^{-4}$ . The batch size is set to 24, and the model is trained for 100 epochs. Early stopping is applied to prevent overfitting. A cosine annealing schedule is adopted to update the learning rate, and gradient clipping with a threshold of 3.0 is applied to stabilize optimization. For both Weibo and PHEME, five-fold cross-validation is adopted to evaluate the model performance. To ensure statistical robustness, each fold is run with three different random seeds, and the reported metrics are averaged across all seeds and folds.

#### 4.4. Performance Comparison

It should be noted that the results of EANN and SAFE in Table 1 are reproduced by us under the strict zero-response constraint, as their original papers did not report evaluations on these datasets under this specific setting. All baselines in the table are therefore compared under identical conditions.

**Table 1.** Performance comparison of content-based methods on the Weibo and PHEME datasets.

| Dataset | Method                 | Accuracy     | Precision    | Recall       | F1           |
|---------|------------------------|--------------|--------------|--------------|--------------|
| Weibo   | EANN [37] <sup>†</sup> | 82.87        | 82.53        | 82.31        | 82.42        |
|         | SAFE [44] <sup>†</sup> | 85.21        | 85.18        | 85.24        | 85.21        |
|         | DMFL [45]              | 90.51        | 89.71        | 90.24        | 89.95        |
|         | SemSim [46]            | 90.97        | 90.92        | 89.83        | 90.31        |
|         | <b>SMD-Net</b>         | <b>92.60</b> | <b>92.23</b> | <b>90.48</b> | <b>91.34</b> |
| PHEME   | EANN [37] <sup>†</sup> | 79.28        | 75.62        | 74.91        | 75.26        |
|         | SAFE [44] <sup>†</sup> | 83.15        | 81.83        | 81.27        | 80.32        |
|         | DMFL [45]              | 88.83        | 87.70        | 88.53        | 87.81        |
|         | SemSim [46]            | 89.94        | 88.77        | 86.56        | 87.61        |
|         | <b>SMD-Net</b>         | <b>90.27</b> | <b>89.48</b> | <b>87.78</b> | <b>88.62</b> |

Bold values indicate the best performance in each column. <sup>†</sup> Results reproduced by us under the strict zero-response setting (source post text and image only). Other results are taken from the corresponding original papers.

Table 1 reports the results under the strict zero-response setting, where both SMD-Net and the compared baselines use only the source post text and its associated image. Under this fair comparison, SMD-Net achieves the best performance on both datasets. On Weibo, SMD-Net reaches 92.60% accuracy and 91.34% F1, exceeding the strongest content-based baseline SemSim by 1.63 and 1.03 percentage points, respectively. On PHEME, SMD-Net achieves 90.27% accuracy and 88.62% F1, improving over SemSim by 0.33 percentage points in accuracy and 1.01 percentage points in F1. These results indicate that the proposed design extracts more discriminative evidence from source-level multimodal content when propagation-based feedback is unavailable.

The margin on Weibo is noticeably larger than that on PHEME. A plausible explanation is that Weibo rumors more often contain stronger suspicious cues in the source post itself, allowing the gains from textual semantic refinement, visual evidence enhancement,

and adaptive fusion to be expressed more directly. By contrast, the smaller but still consistent improvement on PHEME suggests that some cases in this dataset are intrinsically harder to resolve from source content alone.

Table 2 further compares SMD-Net with response-based methods that additionally incorporate user comments, stance signals, or propagation structures. These methods operate under a more informative input setting and are therefore not directly comparable to zero-response models. Nevertheless, SMD-Net remains highly competitive. On Weibo, SMD-Net achieves the best accuracy and F1 among all listed methods, outperforming the strongest response-based baseline GINMCL by 0.80 percentage points in accuracy and 0.94 percentage points in F1. On PHEME, SMD-Net reports the highest accuracy (90.27%), while GINMCL achieves the best F1 (89.00%). A closer look reveals that the recall of SMD-Net (87.78%) on PHEME is lower than that of GINMCL (89.20%) and DEEP (87.90%), suggesting that propagation-aware methods may still help recover certain borderline rumor cases on PHEME. Overall, these results demonstrate that careful modeling of the source post alone can already recover a substantial amount of discriminative information, making SMD-Net a practical alternative for early rumor detection when propagation-based evidence is unavailable.

**Table 2.** Performance comparison of response-based methods on the Weibo and PHEME datasets.

| Dataset | Method         | Accuracy     | Precision    | Recall       | F1           |
|---------|----------------|--------------|--------------|--------------|--------------|
| Weibo   | QSAN [47] ‡    | 71.01        | 71.02        | 67.54        | 67.58        |
|         | EBGCN [48] ‡   | 83.14        | 85.46        | 81.76        | 81.45        |
|         | MFAN [49] ‡    | 88.95        | 88.91        | 88.13        | 88.33        |
|         | CLFFRD [51]    | 91.26        | 90.23        | 89.70        | 89.82        |
|         | GINMCL [50]    | 91.80        | 91.20        | 90.10        | 90.40        |
|         | <b>SMD-Net</b> | <b>92.60</b> | <b>92.23</b> | <b>90.48</b> | <b>91.34</b> |
| PHEME   | QSAN [47] ‡    | 75.13        | 69.97        | 65.80        | 66.87        |
|         | EBGCN [48] ‡   | 82.99        | 81.31        | 79.29        | 79.82        |
|         | MFAN [49] ‡    | 88.73        | 87.07        | 85.61        | 86.16        |
|         | DEEP [52]      | 88.90        | 86.80        | 87.90        | 87.20        |
|         | CLFFRD [51]    | 89.95        | 88.26        | 87.57        | 88.13        |
|         | GINMCL [50]    | 90.00        | <b>90.40</b> | <b>89.20</b> | <b>89.00</b> |
|         | <b>SMD-Net</b> | <b>90.27</b> | 89.48        | 87.78        | 88.62        |

Bold values indicate the best performance in each column. ‡ Results cited from MFAN [49], which evaluated these methods on Weibo and PHEME under comparable settings. Other results are taken from the corresponding original papers.

#### 4.5. Ablation Study

Table 3 summarizes the ablation study results. We examine two aspects: the contribution of each modality and the effect of the proposed modules.

The single-modality results already show that the full model is not simply dominated by one branch. On both datasets, removing either text or image causes a large drop relative to the full model, which means that the zero-response setting still requires genuinely multi-modal evidence. The behavior of the text-only model is particularly revealing. On Weibo, it achieves high recall but much lower precision, suggesting that textual content alone is often enough to detect suspicious or rumor-like semantics, but not enough to filter out false positives reliably. Once visual evidence is added, this imbalance is reduced substantially. A similar gap between single-modality and full-model performance also appears on PHEME.

**Table 3.** Ablation study of SMD-Net on the Weibo and PHEME datasets.

| Method                | Weibo        |              |              |              | PHEME        |              |              |              |
|-----------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                       | Acc          | Prec         | Rec          | F1           | Acc          | Prec         | Rec          | F1           |
| SMD-Net (Text only)   | 78.36        | 72.69        | 79.84        | 76.10        | 75.75        | 69.77        | 77.30        | 73.34        |
| SMD-Net (Image only)  | 73.22        | 70.22        | 65.87        | 67.98        | 70.55        | 67.06        | 62.38        | 64.64        |
| SMD-Net (TransXNet)   | 84.86        | 83.69        | 80.63        | 82.13        | 81.78        | 80.23        | 76.67        | 78.41        |
| SMD-Net (w/o MUDD)    | 89.52        | 88.78        | 86.67        | 87.71        | 86.58        | 85.46        | 83.02        | 84.22        |
| SMD-Net (w/o DA)      | 90.48        | 89.79        | 87.94        | 88.85        | 87.88        | 86.95        | 84.60        | 85.76        |
| SMD-Net (w/o AGF)     | 91.78        | 91.40        | 89.37        | 90.37        | 88.77        | 88.07        | 85.56        | 86.79        |
| <b>SMD-Net (Full)</b> | <b>92.60</b> | <b>92.23</b> | <b>90.48</b> | <b>91.34</b> | <b>90.27</b> | <b>89.48</b> | <b>87.78</b> | <b>88.62</b> |

Bold values indicate the best performance in each column.

TransXNet is adopted as the visual backbone for its ability to preserve broad structural information while retaining sensitivity to local suspicious details, which is consistent with the visual characteristics of rumor-related images.

The module-level ablation also follows a clear pattern. Starting from the TransXNet baseline, each additional component leads to a further improvement on both datasets. On Weibo, accuracy rises step by step from 84.86% to 92.60%, and on PHEME from 81.78% to 90.27%. The increase is not uniform across all metrics, which is actually informative. For example, comparing SMD-Net (w/o AGF) with the full model, the addition of AGF improves recall on PHEME from 85.56% to 87.78%, suggesting that adaptive fusion is particularly helpful for recovering samples that depend on cross-modal interaction cues. At the same time, because the modules are added progressively rather than tested in every possible combination, these results are better interpreted as evidence that the current design works coherently as a whole, rather than as a complete disentangling of all interaction effects.

The ablation study supports the design of the full model. The gains do not come from a single branch or a single module but from combining multimodal evidence, a stronger visual backbone, and a sequence of design choices that remain consistently beneficial under the current setting.

#### 4.6. Efficiency Analysis

Table 4 reports the parameter efficiency and inference latency of the progressive model variants. Starting from the TransXNet+concat baseline (11.54 M parameters, 58.23 ms latency), each added module incurs a moderate increase in both parameters and latency. Adding MUDD introduces 1.30 M additional parameters and 2.48 ms latency overhead, representing an 11.29% parameter increase and a 4.26% latency increase. The further addition of DA brings the total to 12.99 M parameters with 63.25 ms latency, which is the highest latency among the variants but also delivers the strongest detection performance. Finally, incorporating AGF raises the parameter count to 13.14 M, while the latency slightly decreases to 61.94 ms compared to the +MUDD + DA variant. A possible explanation is that the gating mechanism reduces the effective computation in the fusion stage. Overall, the full configuration (+MUDD + DA + AGF) incurs a 13.88% parameter overhead and a 6.37% latency overhead relative to the baseline, which is acceptable given the consistent performance gains observed in the ablation study.

**Table 4.** Parameter efficiency and inference latency comparison across model variants. Baseline results are measured on TransXNet + concat. All latency values are reported in milliseconds (ms) over 30 runs after 5 warmup iterations.  $\Delta$  denotes absolute and relative changes relative to baseline.

| Variant          | Params (M) | $\Delta$ Params | $\Delta$ (%) | Latency (ms)     | $\Delta$ Lat. (ms) | $\Delta$ (%) |
|------------------|------------|-----------------|--------------|------------------|--------------------|--------------|
| Baseline         | 11.54      | —               | —            | $58.23 \pm 1.75$ | —                  | —            |
| +MUDD            | 12.84      | +1.30           | +11.29%      | $60.71 \pm 2.77$ | +2.48              | +4.26%       |
| +MUDD + DA       | 12.99      | +1.45           | +12.53%      | $63.25 \pm 2.84$ | +5.01              | +8.60%       |
| +MUDD + DA + AGF | 13.14      | +1.60           | +13.88%      | $61.94 \pm 2.82$ | +3.71              | +6.37%       |

## 5. Conclusions

This work focused on early multimodal rumor detection in a strict zero-response setting, where prediction must rely on the source post alone. Under this constraint, SMD-Net was designed to strengthen three aspects of the problem at the same time: noisy textual semantics, subtle visual inconsistency cues, and the interaction between the two modalities. The results on Weibo and PHEME show that this design is effective. SMD-Net achieves the strongest overall performance among the compared content-based methods and remains competitive even when placed alongside response-based approaches that use additional propagation-related information. The ablation study further suggests that the final gain comes from the overall design rather than from any single module in isolation.

Several limitations should also be acknowledged. The visual branch is still relatively heavy, and stacking multiple enhancement modules together increases the computational burden. The current model is also limited to text–image inputs, which leaves out other common media forms on social platforms. More importantly, the zero-response formulation has its own boundary. As the PHEME results suggest, some difficult cases are still easier to recover when response or propagation information is available.

Future work will mainly move in two directions: improving efficiency and extending the framework to settings that are closer to real-world social media. In particular, it would be valuable to explore lighter architectures, richer multimodal inputs, and early-detection settings that allow limited contextual evidence while still preserving timeliness.

**Author Contributions:** Conceptualization, Z.Q.; methodology, Z.Q. and X.Z.; software, Z.Q.; validation, Z.Q.; formal analysis, Z.Q.; investigation, Z.Q. and Z.Y.; data curation, Z.Q. and Z.Y.; writing—original draft preparation, Z.Q.; writing—review and editing, X.Z.; visualization, Z.Q.; supervision, X.Z.; project administration, X.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** No new data were created in this study. The original Weibo and PHEME datasets analyzed are publicly available and have been cited in Section 4.1.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|         |                                                                 |
|---------|-----------------------------------------------------------------|
| SMD-Net | Selective and Multiway Differential Perception Enhanced Network |
| DA      | Differential Attention                                          |
| MUDD    | Multiway Dynamic Dense Connection                               |

|     |                                |
|-----|--------------------------------|
| AGF | Adaptive Gated Fusion          |
| GNN | Graph Neural Network           |
| LLM | Large Language Model           |
| RAG | Retrieval-Augmented Generation |
| SSM | Selective State-Space Model    |

## References

1. Allcott, H.; Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **2017**, *31*, 211–236. [\[CrossRef\]](#)
2. Muhammed T, S.; Mathew, S.K. The disaster of misinformation: A review of research in social media. *Int. J. Data Sci. Anal.* **2022**, *13*, 271–285. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Arcuri, M.C.; Gandolfi, G.; Russo, I. Does fake news impact stock returns? Evidence from US and EU stock markets. *J. Econ. Bus.* **2023**, *125*, 106130. [\[CrossRef\]](#)
4. West, D.M.; Kamarck, E.; Kalb, M.; Kaufmann, D.; Chamorro, S.R.A.; Kamarck, J.M.E. *Misinformation Is Eroding the Public's Confidence in Democracy*; Technical Report; Brookings Institution: Washington, DC, USA, 2022.
5. Pattanaik, B.; Mandal, S.; Tripathy, R.M. A survey on rumor detection and prevention in social media using deep learning. *Knowl. Inf. Syst.* **2023**, *65*, 3839–3880. [\[CrossRef\]](#)
6. Sattarov, O.; Choi, J. Detection of rumors and their sources in social networks: A comprehensive survey. *IEEE Trans. Big Data* **2024**, *11*, 1528–1547. [\[CrossRef\]](#)
7. Zhang, L.; Chen, Y.; Luo, X.; Zhu, H.; He, G. Early detection of rumors based on propagation prediction in social media. *Complex Intell. Syst.* **2026**, *12*, 23.
8. Ziari, M.; Charkari, N.M. Rumor detection and propagation on social networks: A survey. *Expert Syst. Appl.* **2026**, *295*, 128798. [\[CrossRef\]](#)
9. Lv, J.; Gao, Y.; Li, L.; Shi, L.; Li, S. Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. *J. King Saud Univ. Comput. Inf. Sci.* **2025**, *37*, 306. [\[CrossRef\]](#)
10. Segura-Bedmar, I.; Alonso-Bartolome, S. Multimodal fake news detection. *Information* **2022**, *13*, 284. [\[CrossRef\]](#)
11. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*; IEEE: Piscataway, NJ, USA, 2019; pp. 39–47.
12. Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; Wei, L. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.* **2021**, *58*, 102610. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Zhang, Q. Rumor Detection Model on Social Media Based on Bi-LSTM and Hierarchical Attention Mechanism. *Hum.-Cent. Intell. Syst.* **2025**, *6*, 35–47. [\[CrossRef\]](#)
14. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2560–2569.
15. Gu, Y.; Ablimit, M.; Hamdulla, A. Fake News Detection Based on Cross-modal Co-attention. In *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*; IEEE: Piscataway, NJ, USA, 2023; pp. 401–406.
16. Tuan, N.M.D.; Minh, P.Q.N. Multimodal fusion with BERT and attention mechanism for fake news detection. In *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
17. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, Philadelphia, PA, USA, 7–9 October 2024.
18. Lou, M.; Zhang, S.; Zhou, H.Y.; Yang, S.; Wu, C.; Yu, Y. TransXNet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 11534–11547. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; Wei, F. Differential transformer. *arXiv* **2024**, arXiv:2410.05258.
20. Xiao, D.; Meng, Q.; Li, S.; Yuan, X. Muddformer: Breaking residual bottlenecks in transformers via multiway dynamic dense connections. *arXiv* **2025**, arXiv:2502.12170.
21. Ma, J.; Chen, C.; Hou, C.; Yuan, X. Kapalm: Knowledge graph enhanced language models for fake news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 3999–4009.
22. Gong, S.; Sinnott, R.O.; Qi, J.; Paris, C.; Nakov, P.; Xie, Z. Multi-Sourced, Multi-Agent Evidence Retrieval for Fact-Checking. *arXiv* **2026**, arXiv:2603.00267.
23. Niu, C.; Guan, Y.; Wu, Y.; Zhu, J.; Song, J.; Zhong, R.; Zhu, K.; Xu, S.; Diao, S.; Zhang, T. VeraCT scan: Retrieval-augmented fake news detection with justifiable reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 266–277.

24. Fetahu, B.; Markert, K.; Anand, A. Fine grained citation span for references in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1990–1999.
25. Castillo, C.; Mendoza, M.; Poblete, B. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*; Association for Computing Machinery: New York, NY, USA, 2011; pp. 675–684.
26. Yang, F.; Liu, Y.; Yu, X.; Yang, M. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 1–7.
27. Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; Shah, S. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1867–1870.
28. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.F.; Cha, M. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2016; pp. 3818–3824.
29. Kumar, S.; Carley, K.M. Tree lstms with convolution units to predict stance and rumor veracity in social media conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 5047–5058.
30. Ma, J.; Gao, W. Debunking rumors on twitter with tree transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5455–5466.
31. Guo, M.; Chen, C.; Hou, C.; Wu, Y.; Yuan, X. FGDGNN: Fine-grained dynamic graph neural network for rumor detection on social media. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 5676–5687.
32. Xu, S.; Liu, X.; Ma, K.; Dong, F.; Riskhan, B.; Xiang, S.; Bing, C. Rumor detection on social media using hierarchically aggregated feature via graph neural networks. *Appl. Intell.* **2023**, *53*, 3136–3149. [[CrossRef](#)]
33. Zeng, Y.; Ding, X.; Cai, B.; Liu, T.; Qin, B. Exploring large language models for effective rumor detection on social media. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 2537–2552.
34. Yu, Z.; Li, X.; Zhang, J. ISMAF: Intrinsic-social modality alignment and fusion for multimodal rumor detection. *Neurocomputing* **2026**, *677*, 133098. [[CrossRef](#)]
35. Qian, S.; Wang, J.; Hu, J.; Fang, Q.; Xu, C. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 153–162.
36. Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; Yu, Y. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1212–1220.
37. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 849–857.
38. Zhang, H.; Qian, S.; Fang, Q.; Xu, C. Multimodal disentangled domain adaption for social media event rumor detection. *IEEE Trans. Multimed.* **2020**, *23*, 4441–4454. [[CrossRef](#)]
39. Xiao, L.; Zhang, Q.; Shi, C.; Wang, S.; Naseem, U.; Hu, L. Msynfd: Multi-hop syntax aware fake news detection. In *Proceedings of the ACM Web Conference 2024*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 4128–4137.
40. Wang, B.; Li, X.; Li, C.; Fu, B.; Pei, S.; Wang, S. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 2304–2314.
41. Dong, Y.; He, D.; Wang, X.; Jin, Y.; Ge, M.; Yang, C.; Jin, D. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2024; Volume 38, pp. 8354–8362.
42. Song, C.; Yang, C.; Chen, H.; Tu, C.; Liu, Z.; Sun, M. CED: Credible early detection of social media rumors. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 3035–3047. [[CrossRef](#)]
43. Zubiaga, A.; Liakata, M.; Procter, R. Exploiting context for rumour detection in social media. In *Social Informatics, Proceedings of the International Conference on Social Informatics*; Springer: Cham, Switzerland, 2017; pp. 109–123.
44. Zhou, X.; Wu, J.; Zafarani, R. SAFE: Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining, Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*; Springer: Cham, Switzerland, 2020; pp. 354–367.

45. Guo, Z.; Liu, H.; Zuo, L.; Wen, J. Bilinear learning with dual-chain feature attention for multimodal rumor detection. *Mathematics* **2025**, *13*, 1731. [[CrossRef](#)]
46. Li, C.; Xu, B.; Wang, M.; He, K. Semantic similarity driven multi-modal model for rumor detection. In *ECAI 2024*; IOS Press: Amsterdam, The Netherlands, 2024; pp. 3749–3756.
47. Tian, L.; Zhang, X.; Wang, Y.; Liu, H. QSAN: A quantum-probability based signed attention network for explainable false information detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1445–1454.
48. Wei, L.; Hu, D.; Zhou, W.; Yue, Z.; Hu, S. Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3845–3854.
49. Zheng, J.; Zhang, X.; Guo, S.; Wang, Q.; Zang, W.; Zhang, Y. MFAN: Multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*; International Joint Conferences on Artificial Intelligence (IJCAI): Palo Alto, CA, USA, 2022; pp. 2413–2419.
50. Deng, L.; Xiao, K.; Rong, H.; Qian, Z. GINMCL: Graph isomorphism network-driven modality enhancement and cross-modal consistency learning for multi-modal fake news detection. *PeerJ Comput. Sci.* **2026**, *12*, e3592. [[CrossRef](#)]
51. Xu, F.; Zeng, L.; Zou, B.; Aw, A.T.; Rong, H. CLFFRD: Curriculum learning and fine-grained fusion for multimodal rumor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 3314–3324.
52. Chen, J.; Liu, L.; Lan, T.; Zhou, F.; Zhang, X. You only query twice: Multimodal rumor detection via evidential evaluation from dual perspectives. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 2570–2583.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.